ABSTRACT OF THE DISCLOSURE

A content-aware application switch and methods thereof intelligently switch client packets to one server among a group of servers in a server farm. The switch uses Layer 7 or application content parsed from a packet to help select the server and to schedule the transmitting of the packet to the server. This enables refined load-balancing and Quality-of-Service control tailored to the application being switched. In another aspect of the invention, a slow-start server selection method assigned an initially boosted server load metric to a server newly added to the group of servers under load balancing. This alleviates the problem of the new server being swamped initially due to a very low load metric compared to that of others. In yet another aspect of the invention, a switching method dependent on Layer 7 content avoids delayed binding in a new TCP session. Layer 7 content is not available during the initial handshaking phase of a new TCP session. The method uses the Layer 7 content from a previous session as an estimate to help select the server and uses a default priority to scheduling the transmitting of the handshaking packets. Updated Layer 7 content available after the handshaking phase is then used to reset the priority for the transmit schedule and becomes available for use in load balancing of the next TCP session.